Regularized Dual Averaging

12M42340　チョウ　シホウ

7 月 3 日

概要

今回は論文 [1] によって提案された Regularized Dual Averaging（正則化付き双対平均法）を紹介する.

1　導入

1.1　Regularized Stochastic Learning Problem

Regularized stochastic learning（正則化付き確率的学習）問題とは

$$\underset{\boldsymbol{w}\in\mathbb{R}^n}{\text{minimize}} \ \{\phi(\boldsymbol{w}) := \mathbb{E}_z f(\boldsymbol{w}, z) + \Psi(\boldsymbol{w})\} \tag{1.1}$$

ただし，$\boldsymbol{z} = (\boldsymbol{x}, y)$ はある未知な分布に従う入力・出力データペア．$f(\boldsymbol{w}, z)$ は $\boldsymbol{w}$ と x を利用して y を予測する損失関数（Loss function）である．なお，次の幾つの仮定を満たされば，最適解を求められる.

- $\Psi(\boldsymbol{w})$ が閉凸関数.
- $\mathbf{dom}\,\Psi = \{\boldsymbol{w} \in \mathbb{R}^n | \Psi(\boldsymbol{w}) < +\infty\}$ が閉集合.
- $f(\boldsymbol{w}, z)$ が凸かつ $\mathbf{dom}\,\Psi$ における劣微分可能.

損失関数 $f(\boldsymbol{w}, z)$ の例：

- Least-squares :　$x \in \mathbb{R}^n, y \in \mathbb{R}$,and $f(\boldsymbol{w}, (\boldsymbol{x}, y)) = \|y - \boldsymbol{w}^\intercal \boldsymbol{x}\|_2^2)$
- Hinge loss:　$x \in \mathbb{R}^n, y \in \{+1, -1\}$,and $f(\boldsymbol{w}, (\boldsymbol{x}, y)) = \max\{0, 1 - y(\boldsymbol{w}^\intercal \boldsymbol{x})\}$
- Logistic regression:　$x \in \mathbb{R}^n, y \in \{+1, -1\}$,and $f(\boldsymbol{w}, (\boldsymbol{x}, y)) = \log(1 + \exp(-y(\boldsymbol{w}^\intercal \boldsymbol{x})))$

正則化項（regularization term）$\Psi(\boldsymbol{w})$ の例：

- ℓ_1-regularization :　$\Psi(\boldsymbol{w}) = \lambda\|\boldsymbol{w}\|, \lambda > 1.$
- ℓ_2-regularization :　$\Psi(\boldsymbol{w}) = \frac{\sigma}{2}\|\boldsymbol{w}\|_2^2, \sigma > 0.$
- Convex Constraints :　$\Psi(\boldsymbol{w}) = \begin{cases} 0 & \boldsymbol{w} \in C \\ +\infty & otherwise \end{cases}$

1.2 一般的なアプローチ

一般的なアプローチは，有限な独立観測値列 $z_1, z_2 \ldots, z_T$ を用いて $\phi(\boldsymbol{w})$ の期待値を近似する．

$$\underset{\boldsymbol{w}\in\mathbb{R}^n}{\text{minimize}} \ \frac{1}{T}\sum_{t=1}^{T} f(\boldsymbol{w}, z_t) + \Psi(\boldsymbol{w}) \tag{1.2}$$

しかし，汎用アルゴリズムを用いて式 (1.2) を直接に解くのは難しい，問題 (1.1) を効率よく解くために Regularized Dual Averaging(RDA) が提案されていた．RDA の本質とは

Regularization + Stochastic Gradient + Dual Averaging + Shrinking

の組み合せアルゴリズムである．

2 Stochastic Gradient Descent(SGD)

2.1 Online Learning

オンライン学習とは，データを 1 つづつ読み込んで、それまでの学習結果を更新すること．正則化なし損失関数 $\min f(\boldsymbol{w}, z)$ に対するのオンライン学習最適化を用いたアルゴリズムは次である．

Algorithm 1 Online Learning Algorithm

initialize: 初期点 $\boldsymbol{w}_1$ を求める．

for $t = 1, 2, 3, \ldots,$ **do**

1. 時刻 t に於ける，入力・出力ペア z_t と $\boldsymbol{w}_t$ を用いて損失関数 $f(\boldsymbol{w}_t, z_t)$ を計算．
2. ステップ 1 の $\boldsymbol{w}_t$ を利用して予測を行い，予測が間違うならば $\boldsymbol{w}_{t+1}$ を更新する．

間違ったら損失が増加する，オンライン学習の目標は累積損失や Regret などの評価関数を最小化すること．

2.2 Regularized online optimization

正則化付きオンライン最適化は，次の Regret 関数

$$R_t(\boldsymbol{w}) := \sum_{\tau=1}^{t}(f_\tau(\boldsymbol{w}_\tau) + \Psi(\boldsymbol{w}_\tau)) - \sum_{\tau=1}^{t}(f_\tau(\boldsymbol{w}) + \Psi(\boldsymbol{w})) \tag{2.1}$$

を使う．

2.3 Stochastic Gradient Descent

次の正則化関数を考え：$\Psi(\boldsymbol{w}) = I_C(\boldsymbol{w}) + \psi(\boldsymbol{w})$，$I_C$ は hard regularization，$\psi(\boldsymbol{w})$ は soft regularization．SGD は

$$\boldsymbol{w}_{t+1} = \Pi_C\big(\boldsymbol{w}_t - \boldsymbol{\alpha}_t(\boldsymbol{g}_t + \boldsymbol{\xi}_t)\big) \tag{2.2}$$

である．ただし，$\boldsymbol{g}_t \in \partial f(\boldsymbol{w}_t, z_t), \boldsymbol{\xi}_t \in \partial\Psi(\boldsymbol{w}_t)$，$\Pi_C(\boldsymbol{w})$ は次に定義される関数である．

$$\Pi_C(\boldsymbol{w}) = \arg\min_{\boldsymbol{w}' \in C} \|\boldsymbol{w} - \boldsymbol{w}'\|_2 \tag{2.3}$$

$\Pi_C(\boldsymbol{w})$ を Euclidean Projection という．

SGD アルゴリズムでは，$\boldsymbol{w}_t$ と $\phi(\boldsymbol{w}_t)$ が反復 t 前の情報 $\{z_1, z_2, \ldots, z_{t-1}\}$ に依存する．(1.1) の最適解 $\boldsymbol{w}^*$ の存在を仮定し，$\phi^* = \phi(\boldsymbol{w}^*)$ とおくと，

$$\lim_{t\to\infty} \phi(\boldsymbol{w}_t) = \phi^*$$

を求めるのは SGD のアプローチである．SGD はオンライン学習版の勾配法である．

SGD は Online Learning の一種として，収束が早い，大量のメモリを必要としない，新たなデータが来たときに，再学習が容易，機械学習にはよく使われて大規模データにも対応できる，しかし，SGD が正則化付き問題に対して特化していない，もっと効率よく解きたい．

3 Regularized Dual Averaging

3.1 Algorithm

Algorithm 2 Regularized dual averaging(RDA) method

input:

- a strongly convex function $h(\boldsymbol{w})$ with modulus 1 on $\mathbf{dom}\,\Psi$,and $\boldsymbol{w}_0 \in \mathbb{R}^n$,such that

$$\boldsymbol{w}_0 = \arg\min_{\boldsymbol{w}} h(\boldsymbol{w}) \in \arg\min_{\boldsymbol{w}} \Psi(\boldsymbol{w}) \tag{3.1}$$

- a pre-detemined nonnegative and nondecreasing sequence β_t for $t \geq 1$

initialize: $\boldsymbol{w}_1 = \boldsymbol{w}_0, \bar{\boldsymbol{g}}_0 = 0$

for $t = 1, 2, 3, \ldots,$ **do**

1. Given the function f_t, compute a subgradient $g_t \in \partial f_t(\boldsymbol{w}_t)$
2. Update the average subgradient $\bar{\boldsymbol{g}}_t$

$$\bar{\boldsymbol{g}}_t = \frac{t-1}{t}\bar{\boldsymbol{g}}_{t-1} + \frac{1}{t}\boldsymbol{g}_t \tag{3.2}$$

3. Compute the next iterate $\boldsymbol{w}_{t+1}$:

$$\boldsymbol{w}_{t+1} = \arg\min_{\boldsymbol{w}} \left\{ \langle \bar{\boldsymbol{g}}_t, \boldsymbol{w} \rangle + \Psi(\boldsymbol{w}) + \frac{\beta_t}{t} h(\boldsymbol{w}) \right\} \tag{3.3}$$

RDA の収束率を保証するために次の不等式を満たす補助関数（auxiliary function）$h(\boldsymbol{w})$ を定義する，

$$h(\alpha\boldsymbol{w} + (1-\alpha)\boldsymbol{u}) \leq \alpha h(\boldsymbol{w}) + (1-\alpha)h(\boldsymbol{u}) - \frac{\sigma}{2}\alpha(1-\alpha)\|\boldsymbol{w} - \boldsymbol{u}\|^2 \tag{3.4}$$

h は **Strongly Convex Function** と呼ばれ，次の性質が成立

$$h(\boldsymbol{w}) \geq h(\boldsymbol{u}) + \langle \boldsymbol{s}, \boldsymbol{w} - \boldsymbol{u} \rangle + \frac{\sigma}{2}\|\boldsymbol{w} - \boldsymbol{u}\|^2, \ \forall \boldsymbol{s} \in \partial h(\boldsymbol{u}) \tag{3.5}$$

$\sigma > 0$ を凸性パラメータ（Convexity parameter）という．$\sigma = 0$ の場合は一般的な凸とは同じ．

RDA の各反復 t では

- ステップ１：従来の勾配法や劣勾配法と同じ，劣勾配を算出する．
- ステップ２：平均劣勾配を更新する．
- ステップ３：RDA で構造した簡単関数を計算し，$\boldsymbol{w}_{t+1}$ を求める．

$\beta_t = \Theta(\sqrt{t})$ の場合，RDA の収束率は

$$\mathbb{E}\phi(\bar{\boldsymbol{w}}_t) - \phi^* \leq \mathcal{O}\Big(\frac{G}{\sqrt{t}}\Big) \tag{3.6}$$

ただし，G は $\boldsymbol{g}_t$ の劣勾配のノルムの上限，$\bar{\boldsymbol{w}}_t = \frac{1}{t}\sum_{\tau=1}^{t} \boldsymbol{w}_t$.

3.2 RDA Methods with General Convex Regularization

$\sigma = 0$ の $\Psi(\boldsymbol{w})$ に対して，$\mathcal{O}(\frac{1}{\sqrt{t}})$ の収束率を手に入れるため，次の正数列 $\{\beta_t\}_{t\geq 1}$ を構造する．

$$\beta_t = \gamma\sqrt{t}, \quad t = 1, 2, 3, \ldots \tag{3.7}$$

1. **Nesterov's dual averaging method**

   $\Psi(\boldsymbol{w})$ を凸集合 C に於ける標示関数に定義する．

   $$\Psi(\boldsymbol{w}) = I_C(\boldsymbol{w}) = \begin{cases} 0 & \boldsymbol{w} \in C \\ +\infty & otherwise \end{cases} \tag{3.8}$$

   また，$h(\boldsymbol{w}) = \frac{1}{2}\|\boldsymbol{w}\|_2^2$ とおくと，式 (3.3) は次に変形できる

   $$\begin{aligned} \boldsymbol{w}_{t+1} &= \arg\min_{\boldsymbol{w}} \Big\{ \langle \bar{\boldsymbol{g}}_t, \boldsymbol{w} \rangle + \Psi(\boldsymbol{w}) + \frac{\beta_t}{t} h(\boldsymbol{w}) \Big\} = \arg\min_{\boldsymbol{w}} \Big\{ \langle \bar{\boldsymbol{g}}_t, \boldsymbol{w} \rangle + I_C(\boldsymbol{w}) + \frac{\gamma}{2\sqrt{t}}\|\boldsymbol{w}\|_2^2 \Big\} \\ &= \Pi_C\Big(-\frac{\sqrt{t}}{\gamma}\bar{\boldsymbol{g}}_t \Big) = \Pi_C\Big(-\frac{1}{\gamma\sqrt{t}} \sum_{\tau=1}^{t} \boldsymbol{g}_\tau \Big) \end{aligned} \tag{3.9}$$

   (3.9) 式が Nesterov(2009) の simple dual averaging とは同じである．近似型（closed form）がないが，効率のよい射影関数が存在する．

   特に，$C = \{\boldsymbol{w} \in \mathbb{R}^n \mid \|\boldsymbol{w}\|_1 \leq \delta\}, \delta > 0$ ならば，正則化項は hard ℓ_1-regularization になる．

2. **Soft ℓ_1-regularization**

   $\Psi(\boldsymbol{w}) = \lambda\|\boldsymbol{w}\|_1, \lambda > 0$ と $h(\boldsymbol{w}) = \frac{1}{2}\|\boldsymbol{w}\|_2^2$ とおくと，次の

   $$\boldsymbol{w}_{t+1}^{(i)} = \begin{cases} 0 & \text{if} \|\bar{\boldsymbol{g}}_t^{(i)}\| \leq \lambda, \\ -\frac{\sqrt{t}}{\gamma}\big(\bar{\boldsymbol{g}}_t^{(i)} - \lambda \operatorname{sign}(\bar{\boldsymbol{g}}_t^{(i)})\big) & \text{otherwise}, \end{cases} \quad i = 1, 2, \ldots, n \tag{3.10}$$

   ただし，$\boldsymbol{w} > 0$ ならば $\operatorname{sign}(\boldsymbol{w}) = 1$，$\boldsymbol{w} \leq 0$ ならば $\operatorname{sign}(\boldsymbol{w}) = 0$,

3.3 RDA Methods with Strongly Convex Regularization

もし $\Psi(\boldsymbol{w})$ は $\sigma > 0$ の Strongly Convex Function ならば，ある収束率を $\mathcal{O}(\ln t)$ に超えないような非負非減少数列 $\{\beta_t\}_{t\geq 1}$ を構造し，元問題が $\mathcal{O}(\frac{\ln t}{t})$ のスピードで収束できる．簡単にするため，$\beta_t = 0$ に設定し，$h(\boldsymbol{w})$ もいらなくなる．

1. **Mixed ℓ_1/ℓ_2^2-regularization**
   $\Psi(\boldsymbol{w}) = \lambda\|\boldsymbol{w}\|_1 + \frac{\sigma}{2}\|\boldsymbol{w}\|_2^2, \lambda > 0, \sigma > 0$ とおくと

$$\boldsymbol{w}_{t+1}^{(i)} = \begin{cases} 0 & \text{if } \|\bar{\boldsymbol{g}}_t^{(i)}\| \leq \lambda, \\ -\frac{1}{\sigma}\left(\bar{\boldsymbol{g}}_t^{(i)} - \lambda \operatorname{sign}\left(\bar{\boldsymbol{g}}_t^{(i)}\right)\right) & \text{otherwise} \end{cases} \quad i = 1, 2, \ldots, n \tag{3.11}$$

2. **Kullback-Leibler(KL) divergence regularization** 略

4 Regret bounds and convergence rates

次は式 (2.1) に定義された関数 $R_t(\boldsymbol{w})$ の上界下界を求める．便宜上で次の記号を定義する．

$$\Delta_t := (\beta_0 - \beta_1)h(\boldsymbol{w}_2) + \beta_t D^2 + \frac{L^2}{2}\sum_{\tau=0}^{t-1}\frac{1}{\sigma\tau + \beta_\tau}, \quad t = 1, 2, 3, \ldots \tag{4.1}$$

D と L はある定数，σ は $\Psi(\boldsymbol{w})$ の凸性パラメータ，$\{\beta_\tau\}_{\tau=1}^t$ は RDA の非負非減少入力数列である．また，収束分析を簡単にするために $\tau = 0$ の時にも値をとれるような $\beta_0 > 0$ となる β_0 を加える．実は $\beta_0 = \beta_1$ とおけば，(4.1) の $(\beta_0 - \beta_1)h(\boldsymbol{w}_2)$ を消せる．また，$t = 1$ における $\boldsymbol{w}_2$ が分かるため，Δ_1 も求められる．すべての $D > 0$ に対して，

$$\mathcal{F}_D := \left\{\boldsymbol{w} \in \mathbf{dom}\,\Psi \;\middle|\; h(\boldsymbol{w}) \leq D^2\right\} \tag{4.2}$$

に定める．

定理 4.1. *Algorithm 2* で $\{\boldsymbol{w}_\tau\}_{\tau=1}^t$ と $\{\boldsymbol{g}_\tau\}_{\tau=1}^t$ を生成し，$\exists L$，$\forall t \geq 1, \boldsymbol{g}_t$ の双対ノルム $\|\boldsymbol{g}_t\|_* \leq L$，が成り立つならば，

$$\forall t \geq 1, \ \forall \boldsymbol{w} \in \mathcal{F}_D, \ R_t(\boldsymbol{w}) \leq \Delta_t \tag{4.3}$$

もし正則化関数 $\Psi(\boldsymbol{w})$ の凸性パラメータ $\sigma = 0$ ならば，$\beta_0 = \beta_1$ とおき，(3.7) の $\{\beta_t\}_{t\geq 1}$ を用いて次の Δ_t を導ける．

$$\Delta_t = \gamma\sqrt{t}D^2 + \frac{L^2}{2\gamma}\left(1 + \sum_{\tau=1}^{t-1}\frac{1}{\sqrt{\tau}}\right) \leq \gamma\sqrt{t}D^2 + \frac{L^2}{2\gamma}\left(1 + (2\sqrt{t} - 2)\right) \leq \left(\gamma D^2 + \frac{L^2}{\gamma}\right)\sqrt{t} \tag{4.4}$$

Δ_t を最小化にする γ は $\gamma^* = \frac{L}{D}$，従って，

$$R_t(\boldsymbol{w}) \leq 2LD\sqrt{t} \tag{4.5}$$

正則化関数 $\Psi(\boldsymbol{w})$ の $\sigma > 0$ の場合，$h(\boldsymbol{w}) = \frac{1}{\sigma}\Psi(\boldsymbol{w})$ にし，β_t によって違う Δ_t を求められる．

1. **Positive constant sequences**
   $\beta_t = \sigma, t \geq 1, \beta_0 = \beta_1$ に対し

$$\Delta_t = \sigma D^2 + \frac{L^2}{2\sigma}\sum_{\tau=1}^{t}\frac{1}{\tau} \leq \sigma D^2 + \frac{L^2}{2\sigma}(1+\ln t) \tag{4.6}$$

2. **The logrithmic sequence**
   $\beta_t = \sigma(1+\ln t), t \geq 1, \beta_0 = \sigma$ に対し

$$\Delta_t = \sigma(1+\ln t)D^2 + \frac{L^2}{2\sigma}\Big(1+\sum_{\tau=1}^{t-1}\frac{1}{\tau+1+\ln\tau}\Big) \leq \Big(\sigma D^2 + \frac{L^2}{2\sigma}\Big)(1+\ln t) \tag{4.7}$$

3. **The zero sequence**
   $\beta_t = 0, t \geq 1, \beta_0 = \sigma, h(\boldsymbol{w}) = \frac{1}{\sigma}\Psi(\boldsymbol{w})$ に対し

$$\Delta_t \leq \Psi(\boldsymbol{w}_2) + \frac{L^2}{2\sigma}\Big(1+\sum_{\tau=1}^{t}\frac{1}{\tau}\Big) \leq \frac{L^2}{2\sigma}(6+\ln t) \tag{4.8}$$

   定理 4.2. もし問題 (1.1) には最適解 $\boldsymbol{w}^*$ が存在し, $\exists D > 0, h(\boldsymbol{w}^*) \leq D^2$ と $\exists L > 0, \mathbb{E}\|\boldsymbol{g}\|_*^2 \leq L^2, \forall \boldsymbol{g} \in \partial f(\boldsymbol{w}, z), \boldsymbol{w} \in \mathbf{dom}\,\Psi$ を満たす, すべて $t \geq 1$ に対して,

$$\boldsymbol{E}_\phi(\bar{\boldsymbol{w}}_t) - \phi(\boldsymbol{w}^*) \leq \frac{\Delta_t}{t}, \ \ \textit{where}\ \ \bar{\boldsymbol{w}}_t = \frac{1}{t}\sum_{\tau=1}^{t}\boldsymbol{w}_\tau \tag{4.9}$$

参考文献

[1] L. Xiao, "Dual averaging methods for regularized stochastic learning and online optimization," Journal of Machine Learning Research, vol. 11, pp. 2543-2596, Dec. 2010. [Online]. Available: `http://portal.acm.org/citation.cfm?id=1756006.1953017`

[2] Yu. Nesterov, Primal-dual subgradient methods for convex problems. Mathematical Programming,120(1):221-259, 2009.

[3] S. Shalev-Shwartz, O. Shamir, N. Srebro, and K. Sridharan, "Stochastic convex optimization," in Proceedings of the Conference on Learning Theory (COLT), 2009. [Online]. Available: `http://eprints.pascal-network.org/archive/00005408/`

[4] Y. Nesterov, Introductory Lectures on Convex Optimization: A Basic Course (Applied Optimization), 1st ed. Springer Netherlands.

[5] Steve Wright, NIPS Tutorial, 6 December 2010, `http://pages.cs.wisc.edu/~swright/nips2010/sjw-nips10.pdf`